TODD WARCZAK

Data Scientist

CONTACT

twarczak@gmail.com
(206) 999-6478
White River Junction, VT
toddwarczak.netlify.app
LinkedIn
Github
@TWarczak

EDUCATION

Ph.D.
Molecular & Cellular Biology
Dartmouth College
September 2012 - August 2020
Hanover, NH

B.S.
Biology
University of Utah
September 2007 - May 2012
Salt Lake City, UT

SKILLS

R ★★★★
Python ★★
AWS SageMaker/S3/EC2/...
SQL
Github
Markdown/Jupyter Notebooks
Data Cleaning/Wrangling
Data Visualization
Probabilities/Statistics
Time-Series Forecasting
Regression/Classification
TensorFlow/Tidymodels/H2O...

WORK EXPERIENCE

Molecular Biologist

Dartmouth College

September 2012 - September 2020 / Hanover, NH

- Engineered genome-wide association study (GWAS) to identify genes controlling arsenic tolerance in plants. Determined Arabidopsis gene NIP1;1 on 4th chromosome as the major genetic factor for tolerating arsenic. All data cleaned/wrangled/analyzed in R.
- Built lab RNA-seq pipeline for gene expression of 25000+ plant genes with R scripts for unsupervised learning (PCA, hierarchical clustering), regression (GLMs/ANOVA), exploratory data analysis, and statistical tests.
- Developed first cell-type specific expression maps for plant genes involved in root arsenic acquisition, efflux, and sequestration (using R).

PROJECTS

Predicting Churn using AWS SageMaker & Local RStudio

- Utilized SageMaker and built-in XGBoost algorithm to train, tune, evaluate, and deploy model for predicting bank customer churn in the SLICED season 1 episode 7 Kaggle competition. (Blog, Github)
- Configured local RStudio to make API calls to SageMaker using SageMaker Python SDK and {reticulate} R package.
- Best performing model deployed as SageMaker endpoint for real-time predictions on holdout data (w/ minimal feature engineering and pre-processing). Predictions submitted to SLICED competition received a score of 0.07622 (LogLoss), which would have placed 8th out of 130 entries.

Forecasting Daily Sales with {modeltime}

- Forecasted 3 months of daily sales utilizing the {modeltime} package in R to help 'Superstore' company selling furniture, technology, and office supplies manage supply-chain/inventory decisions in Q1 (data: Kaggle).
- Exploratory data analysis and 11 models tested. 6 {tidymodels} workflows tuned for individual forecasts plus one weighted ensemble (Support Vector Machine/Neural Network-AR/Random Forest/Prophet-XGboost) forecast. (Blog, Github)

TidyTuesday

- Weekly twitter project focusing on cleaning, wrangling, summarizing, and arranging a new dataset in R to produce a single chart. Typically using {ggplot2} and {tidyverse} tools. Shared via #TidyTuesday. (Github)